II Série • N.º 5 • Junho 2008

Revista Portuguesa de Cirurgia

II Série • N.º 5 • Junho 2008

Revista Portuguesa de Cirurgia

ISSN 1646-6918

Órgão Oficial da Sociedade Portuguesa de Cirurgia

Inferência estatística: intervalos de confiança vs. testes de hipóteses

Nuno Lunet, Milton Severo, Henrique Barros

Serviço de Higiene e Epidemiologia
Faculdade de Medicina da Universidade do Porto

INFERÊNCIA ESTATÍSTICA

A análise estatística de estudos realizados no âmbito da investigação biomédica baseia-se em observações efectuadas em amostras, a partir das quais se procura inferir, ou seja, generalizar para as populações[1] de onde estas foram extraídas, em particular, e mais globalmente para todas as que partilhem características semelhantes. A inferência estatística consiste assim em encontrar os processos adequados para passar do particular para o geral, extrapolar de uma amostra para a população [1].

Vale a pena, no entanto, lembrar que a natureza dos elementos humanos ou dos indivíduos ou parcelas vivas com que trabalhamos é, por essência, mutável e daí que o conceitos estatísticos tradicionais de população e amostra, uma um subconjunto da outra, não se devam em bom rigor usar: não há populações mas apenas amostras pois que toda a gente do mundo hoje avaliada não passa de uma amostra da experiência de vida amanhã diferente. Mas optemos por uma aproximação mais tradicional da estatística não especificamente aplicada a pessoas.

A avaliação de todos os elementos que compõem uma população permite calcular qualquer parâmetro [2] da população (*e.g.*: valor médio da colesterolemia na população; proporção de fumadores na população) mas, de um modo geral, não é possível, nem razoável, avaliar as populações na sua totalidade. Observam-se habitualmente subconjuntos das populações – amostras – que contêm informação sobre os parâmetros da população, permitindo estimá-los. As estatísticas [3] obtidas pela avaliação de cada amostra (*e.g.*: valor médio da colesterolemia; proporção de fumadores) são estimativas imprecisas dos parâmetros da população, o que se reflecte no facto de diferentes amostras originarem estatísticas que podem também ser diferentes. Na figura 1 é apresentado um exemplo em que o valor médio da colesterolemia da população é 1,37 g/L. Extraindo, por exemplo, 100 amostras de 100 indivíduos, a média da colesterolemia em cada amostra variou entre 1,16 g/L e 1,57 g/L, ilustrando o quão imprecisas podem ser as estimativas fornecidas por diferentes amostras da mesma população.

[1] O termo população é aqui utilizado como sinónimo de um conjunto de pessoas, animais, vegetais, acontecimentos, ou quaisquer observações, que frequentemente não é possível estudar na sua totalidade, designadamente por não ser possível enumerar todos os seus elementos, ou pelos elevados custos e dificuldades logísticas associadas ao estudo de populações de grandes dimensões [2].

[2] Um parâmetro é um valor calculado a partir de uma população utilizando todos os seus elementos, pelo que não está associada variação aleatória a este valor. Em populações de grandes dimensões, como as com que frequentemente trabalhamos, é difícil ou mesmo impossível calcular um qualquer parâmetro [1].

[3] Uma estatística é uma quantidade calculada a partir de um conjunto de observações de uma amostra, isto é, medidas sumárias derivadas de amostras (*e.g.*: média amostral, desvio padrão amostral, proporção amostral) [1].

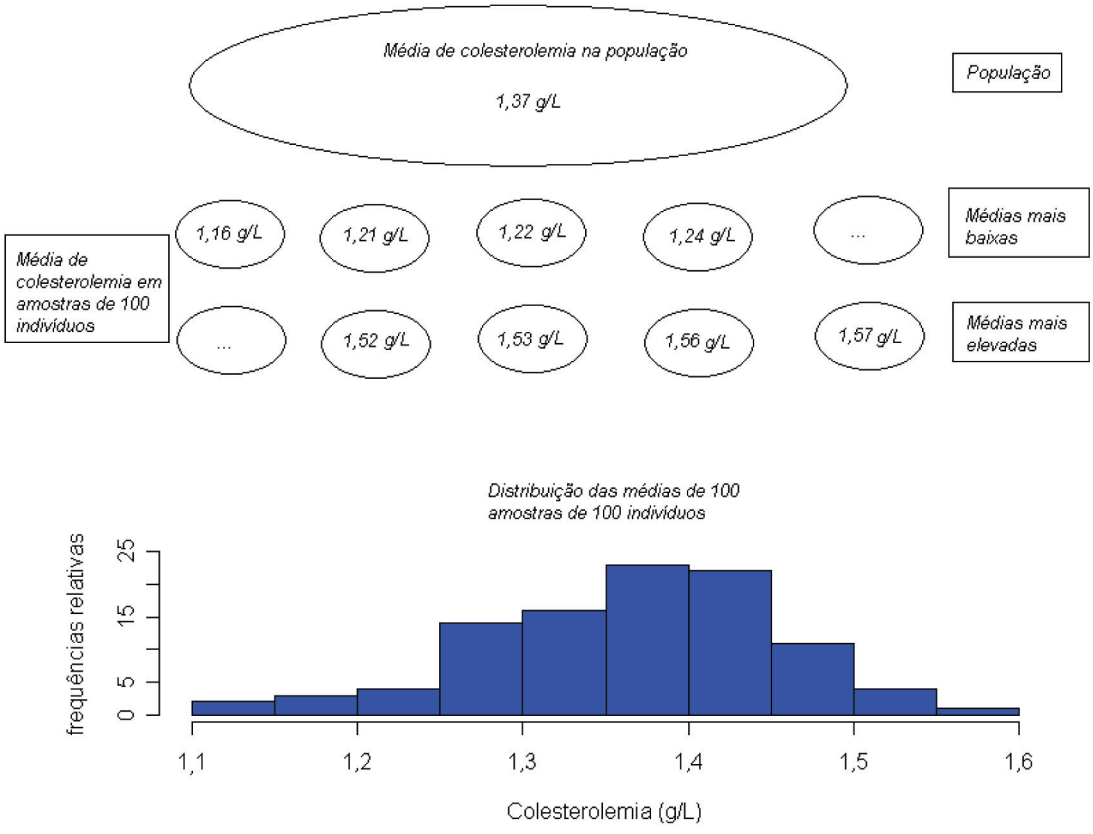

Figura 1. **Distribuição dos valores médios de colesterolemia observados em 100 amostras (cada uma com 100 indivíduos) extraídas de uma população com média de 1,37 g/L.**

A generalização para a população das observações efectuadas em amostras – inferência estatística – implica, para além da avaliação de amostras representativas da população, a quantificação da imprecisão associada às estimativas obtidas com uma amostra (estimação com intervalos de confiança) ou quantificação da probabilidade da amostra provir de uma população com um determinado parâmetro (teste de hipóteses). Este artigo pretende descrever os aspectos essenciais destas duas abordagens, colocando em evidência as suas diferenças e pontos comuns.

ESTIMAÇÃO COM INTERVALOS DE CONFIANÇA

Diferentes amostras da mesma população tendem a originar estimativas diferentes, mas cada uma pode ser vista como uma estimativa pontual do parâmetro populacional que se pretende. A imprecisão de cada estimativa pode ser também calculada com base na informação disponível na amostra: a partir do desvio padrão da amostra (medida da dispersão das observações numa amostra) pode avaliar-se a dispersão das estimativas obtidas de várias amostras diferentes, o chamado erro padrão.

Para uma grande parte das estatísticas (*e.g.*: médias, proporções, taxas, coeficientes de regressão) pode efectuar-se o cálculo do intervalo de confiança com base no mesmo princípio geral, adicionando ou subtraindo à estimativa pontual um múltiplo do respectivo erro padrão [3] (Fórmula 1).

$$[Estimativa - z_{1-\alpha/2} \times Erro\ Padrão;\ Estimativa + z_{1-\alpha/2} \times Erro\ Padrão] \quad (1)$$

Um determinado intervalo deverá conter uma proporção correspondente da distribuição da estatística de interesse. Essa proporção, 1-α, é designada nível de confiança, sendo que α se denomina nível de significância.

A distribuição das estatísticas assume-se Normal, de acordo com o Teorema do Limite Central[4], sendo pos-

sível converter a distribuição normal padronizada na distribuição normal correspondente à estatística em estudo multiplicando-a pelo erro padrão da estimativa e adicionando-lhe o valor da estimativa.

Para o cálculo do intervalo de confiança, multiplica-se o erro padrão por cada um dos valores correspondentes aos valores da distribuição normal padronizada para os quais a probabilidade de se observarem valores inferiores é $\alpha/2$, $z_{\alpha/2}$, e a probabilidade de se observarem valores superiores corresponde a $-1\alpha/2$, $z_{1-\alpha/2}$. A probabilidade correspondente ao intervalo entre estes dois valores é $1-\alpha/2-\alpha/2=1-\alpha$. Como a distribuição Normal padronizada é simétrica em torno de zero $z_{\alpha/2} = -z_{1-\alpha/2}$.

Para um nível de confiança a 95% o nível de significância é de 5% (α=00,5), o $z_{0,0975}$ (valor tabelado da distribuição Normal Padronizada) é aproximadamente 1,96. Para um nível de significância de 1%, o valor da distribuição Normal Padronizada que se multiplica pelo erro padrão para o cálculo do intervalo de confiança é mais elevado ($z_{0,0995}$=2,58) e a amplitude do intervalo de confiança é também maior, reflectindo o aumento do nível de confiança. O contrário observa-se quando o nível de significância é de 10% ($z_{0,95}$=1,64).

Exemplo:

Foi avaliada uma amostra representativa da população do Porto (n=200) com o objectivo de determinar valor médio (μ) da colesterolemia dos residentes na cidade, tendo-se obtido uma estimativa, $\bar{x}$, da média, μ, da colesterolemia de 1,37 (g/L). Sabendo que a variável colesterolemia segue uma distribuição aproximadamente normal, para calcular o intervalo de confiança para μ é necessário conhecer $\bar{x}$, o desvio padrão da população (σ) e a dimensão da amostra (n). Assumindo que o desvio padrão da população é igual ao desvio padrão amostral, s, sendo neste caso 0,42. Temos então que o intervalo de confiança a 95% é dado por:

$$\left[\bar{x} - z_{1-\alpha/2} \times \frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \times \frac{s}{\sqrt{n}}\right] =$$

$$= \left[1{,}37 - 1{,}96 \times \frac{0{,}42}{\sqrt{200}}; 1{,}37 + 1{,}96 \times \frac{0{,}42}{\sqrt{200}}\right] = [1{,}31;\ 1{,}43]$$

Para um nível de confiança de 99% o intervalo de confiança é:

$$\left[1{,}37 - 2{,}58 \times \frac{0{,}42}{\sqrt{200}}; 1{,}37 + 2{,}58 \times \frac{0{,}42}{\sqrt{200}}\right] = [1{,}29;\ 1{,}45]$$

A amplitude de um intervalo de confiança depende do tamanho da amostra, da variabilidade da característica estudada e do nível de confiança requerido, como ilustrado na figura 2. Mantendo os outros dois factores constantes a precisão das estimativas aumenta (menor amplitude dos intervalos de confiança) com o tamanho da amostra e a amplitude dos intervalos de confiança aumenta quando se opta por um nível de confiança mais elevado (maior probabilidade do intervalo de confiança conter o parâmetro que se pretende estimar). Os intervalos de confiança têm menor amplitude (a precisão das estimativas é maior) quando a variabilidade da característica estudada é menor, mas este aspecto não é controlado pelo investigador.

Como se deve interpretar um intervalo de confiança?

O significado de um intervalo de confiança pode ser facilmente compreendido se considerar uma situação teórica em que um estudo é repetido um grande número de vezes, utilizando amostras diferentes extraídas da mesma população. Para cada amostra é obtida uma estimativa pontual e pode calcular-se um intervalo de confiança, e cada um desses intervalos de confiança poderá, ou não, conter o parâmetro populacional que se pretende estimar. Espera-se que a proporção de intervalos de confiança que irá conter o parâmetro populacional seja a correspondente ao nível de confiança para o qual os intervalos foram calculados. Para intervalos de confiança a 95%, esperamos que apenas

[4] Segundo o Teorema do Limite Central, qualquer que seja a distribuição de uma variável, as médias de várias amostras aleatórias seguem uma distribuição normal, desde que a amostra seja suficientemente grande [1].

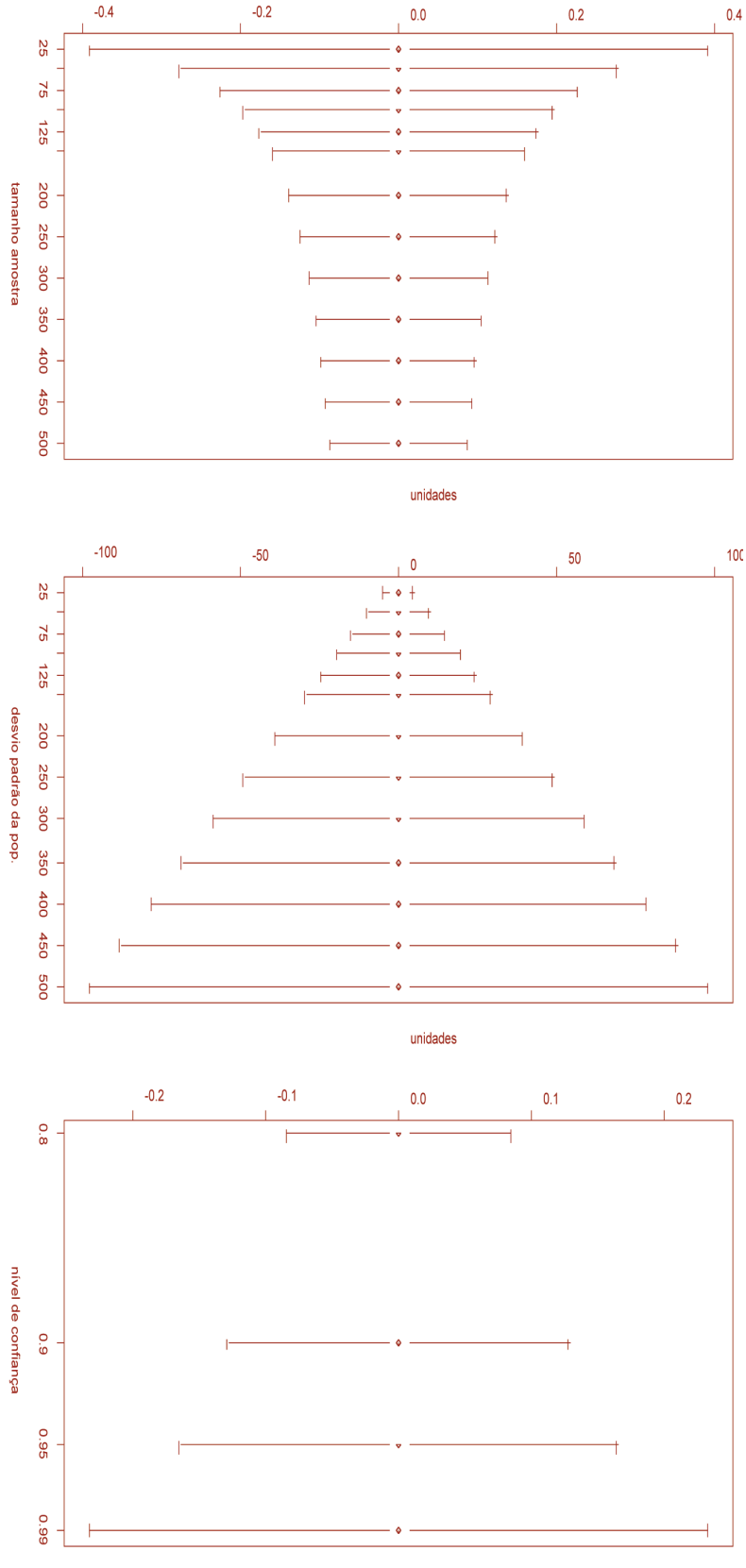

Figura 2. **Variação da amplitude de um intervalo de confiança com o tamanho e a variabilidade de amostra, e com o nível de confiança**

5% dos intervalos calculados não contenham o valor do parâmetro populacional que se desconhece. A figura 3 ilustra esta situação para uma simulação com 100 amostras, podendo verificar-se que neste caso o valor real da média populacional não está contido em 5 dos intervalos. Este valor tende para 5 com o aumento do número de amostras, mas não é necessariamente 5.

Pode esperar-se que os intervalos de confiança a 95% não contenham o parâmetro populacional em apenas 5% das vezes que são calculados, mas perante uma única amostra, como habitualmente acontece, não nos é possível saber se o intervalo calculado contém, ou não, o valor correspondente ao parâmetro populacional.

TESTES DE HIPÓTESES

Os testes de hipóteses são frequentemente utilizados para inferência estatística, com o objectivo de verificar se as estimativas amostrais são compatíveis com determinados parâmetros da população.

A realização de um teste de hipóteses inicia-se com a definição de uma hipótese nula (H_0), que habitualmente consiste na rejeição da hipótese de investigação [5]. Para cada hipótese nula é definida uma hipótese alternativa (H_1), compreendendo resultados não previstos na hipótese nula.

Exemplo:

Numa investigação com o objectivo de confirmar que o valor médio de colesterol sérico nos homens é diferente do observado

[5] Em algumas situações a hipótese nula não consiste na rejeição da hipótese de investigação. Por exemplo, nos testes de hipóteses para a normalidade de uma distribuição, a hipótese de investigação corresponde a H_0

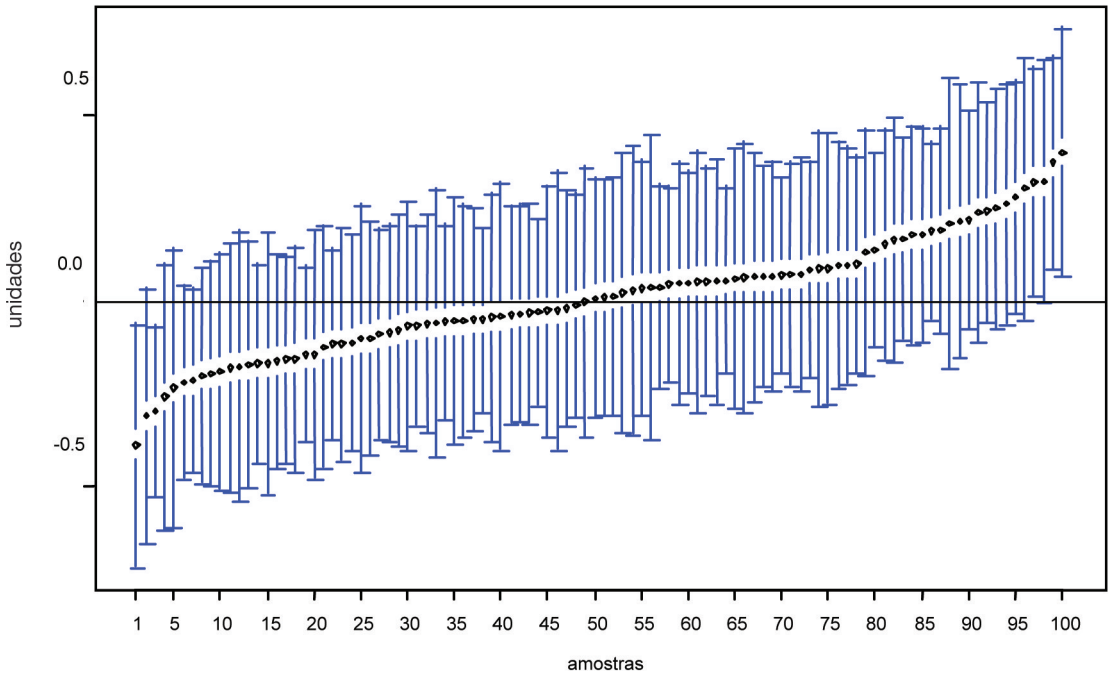

Figura 3. **Intervalos de confiança a 95% correspondentes a 100 amostras extraídas de uma população com distribuição normal, média igual a 0 e variância igual a 1 (simulação).**

nas mulheres, a hipótese nula seria a de que não há diferenças entre as populações de que foram extraídas respectivamente as amostras de mulheres e de homens relativamente ao valor médio de colesterol sérico (H_0: colesterolemia média na população de mulheres = colesterolemia média na população de homens), o que é equivalente a testar a hipótese de que a diferença entre as médias de colesterolemia nas populações das mulheres e dos homens é igual a zero (H_0: Diferença entre a média de colesterolemia nas mulheres e nos homens = 0). As hipóteses alternativas seriam respectivamente: H_1: Colesterolemia média na população de mulheres ≠ Colesterolemia média na população de homens, ou, equivalentemente, H_1: Diferença entre a média de colesterolemia nas mulheres e nos homens ≠ 0).

De seguida é calculada a probabilidade de, numa amostra extraída da população para a qual se pretende inferir, se obter a estimativa observada ou valores ainda mais extremos (mais afastados do parâmetro considerado na hipótese nula), simplesmente devido ao acaso, pela variação que se espera que ocorra de amostra para amostra, se a hipótese nula for verdadeira. Esta probabilidade é habitualmente designada por valor de prova (P).

Estatísticas de teste e o valor de prova (P)

Em geral, para avaliar o afastamento de uma estimativa do parâmetro considerado na Hipótese nula divide-se a diferença entre a estimativa e o parâmetro pelo erro padrão. Ou seja, o afastamento da estimativa relativamente ao parâmetro é expresso em múltiplos do erro padrão (dispersão das estimativas fornecidas por diferentes amostras da mesma população), tomando a designação de estatística de teste (Fórmula 2) [4].

$$Estatística\ de\ Teste = \left| \frac{Estatística - Parâmetro}{EP} \right| \qquad (2)$$

A estatística de teste tem uma determinada distribuição quando a hipótese nula é verdadeira, frequentemente Normal, pelo Teorema do Limite Central. Deste modo, é possível calcular a probabilidade (P) de se observarem valores iguais aos obtidos na amostra em estudo, ou mais extremos (menos prováveis) no caso da Hipótese nula ser verdadeira, correspondendo o valor de prova às caudas da distribuição da estatística de teste (Figura 4).

Qualquer que seja o teste estatístico utilizado, o resultado traduz-se sempre na rejeição ou não rejeição

da Hipótese nula que é testada, comparando o valor de prova com um ponto de corte previamente estabelecido, o nível de significância, . Se a probabilidade for inferior a α, a Hipótese nula é rejeitada, e o resultado é considerado estatisticamente significativo. Para valores de P superiores ou iguais a α, a Hipótese nula não é rejeitada e o resultado é designado não significativo.

Considerando que a distribuição da estatística de teste é a Normal Padronizada, a Hipótese nula é rejeitada caso a estatística de teste seja superior a $z_{1-\alpha/2}$ ou inferior a $z_{\alpha/2}=-z_{1-\alpha/2}$ (figura 4).

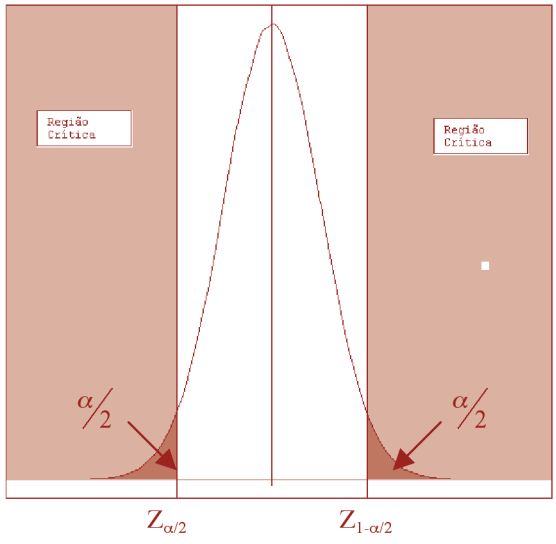

Figura 4. **Região Crítica (área sombreada) para um teste bilateral.**

O valor de Prova corresponde à probabilidade de, quando a Hipótese nula é verdadeira, a estatística de teste tomar um valor igual ou mais extremo do que aquele que é observado, indicando em que medida os dados contradizem a hipótese nula. É comum considerar-se que para valores de prova maiores que 0,05 não existem evidências estatísticas contra H_0, para valores menores que 0,05 existe um evidência moderada contra a H_0, para valores menores que 0,01 existe uma evidência forte, e para valores menores que 0,001 existe uma evidência muito forte [5]. Contudo, a fixação de um nível de significância é francamente arbitrária na medida em que os dados podem contradizer a hipótese nula em maior ou menor grau, e tem-se observado uma tendência para que se apresente o valor de P (e.g.: P=0,042 ou P=0,175), em vez de P<0,05 ou P>0,05, correspondendo a uma quantificação da medida em os dados contradizem a hipótese nula, mas a sua interpretação passa pela comparação do valor obtido com o nível de significância considerado [4].

Exemplo:

Numa investigação com o objectivo de confirmar que existem diferenças no valor médio da colesterolemia entre homens e mulheres, podem definir-se as seguintes hipóteses

$$H_0: \mu_{mulheres}=\mu_{homens} \text{ vs. } H_1: \mu_{mulheres} \mu\neq\mu_{homens},$$

ou equivalentemente,

$$H_0: \mu_{mulheres}-\mu_{homens} =0 \text{ vs. } H_1: \mu_{mulheres}-\mu_{homens} \neq=0$$

Utilizando a formula 2, a estatística de teste que é utilizada para verificar a plausibilidade da H_0 é seguinte:

$$ET = \left| \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \right| > z_{1-\alpha/2}$$ [6]

É então possível construir a seguinte regra de decisão para um nível significância de 5%, se a Estatística de Teste for superior a $z_{1-\alpha/2}$=1,96 ou inferior a $-z_{1-\alpha/2}$=-1,96 rejeita-se a H_0. A região de rejeição ou região critica é região indicada a cinzento na figura 5.

No sexo feminino: $\bar{x}_1$=1,37, s_1=0,42 e n_1=110; no sexo masculino: $\bar{x}_2$=1,38, s_2=0,43 e n_2=90. Logo

$$ET = \left| \frac{(1{,}37 - 1{,}38) - 0}{\sqrt{0{,}42^2/110 + 0{,}43^2/90}} \right| = 0{,}16$$, ou seja, considerando a regra de decisão, não se rejeita H_0, ou seja, a amostra é compatível com uma população em que média de colesterolemia é igual em ambos os sexos.

[6] O erro padrão da diferença de duas médias é igual a:

$$\sqrt{s_1^2/n_1 + s_2^2/n_2}$$

Contudo, a utilização do valor de P é bem mais informativa.

considerando o valor da estatística obtido neste exemplo (0,16), o cálculo do valor de P significa permite conhecer a probabilidade de obter valores iguais ou superiores 0,16 e valores iguais ou inferiores -0,16. O cálculo do valor de P obtido com qualquer software estatístico ou em tabelas da distribuição das estatísticas teste. Neste caso, P foi 0,873.

Testes bilaterais e unilaterais

Os testes de hipóteses são por vezes descritos como bilaterais (*two-tailed*), e se não houver especificação assumem-se como tal. Um teste bilateral calcula a probabilidade de afastamento da hipótese nula em qualquer das direcções (a Hipótese alternativa compreende os valores do parâmetro tanto inferiores como superiores aos afirmados na Hipótese nula repartindo-se a região crítica para a rejeição da hipótese nula pelos dois extremos da distribuição (figura 5, gráfico central). Deste modo, os testes bilaterais são mais conservadores do que os unilaterais, mas não é necessariamente esse o motivo pelo qual são utilizados preferencialmente. Na verdade, é razoável assumir os desvios relativamente à hipótese nula possam ocorrer em duas direcções. Por exemplo, na comparação da colesterolemia média entre homens e mulheres é legítimo admitir que o valor médio possa ser mais alto ou mais baixo em qualquer dos sexos. A opção por um teste unilateral implicaria assumir que apenas seria de esperar que os homens pudessem ter valores mais altos que os das mulheres ou, em alternativa, o inverso.

A opção por um teste unilateral deverá ser sempre efectuada antes da análise dos dados, com base na natureza das relações que se pretendem investigar. Os poucos estudos publicados que utilizam testes unilaterais apresentam valores de P entre 0,025 e 0,05, sugerindo que a opção por testes unilaterais em detrimento dos bilaterais terá sido motivada pela maior facilidade em obter um resultado estatisticamente significativo quando os valores de P são próximos de 0,05 [2]. O *Journal of the National Cancer Institute*, por

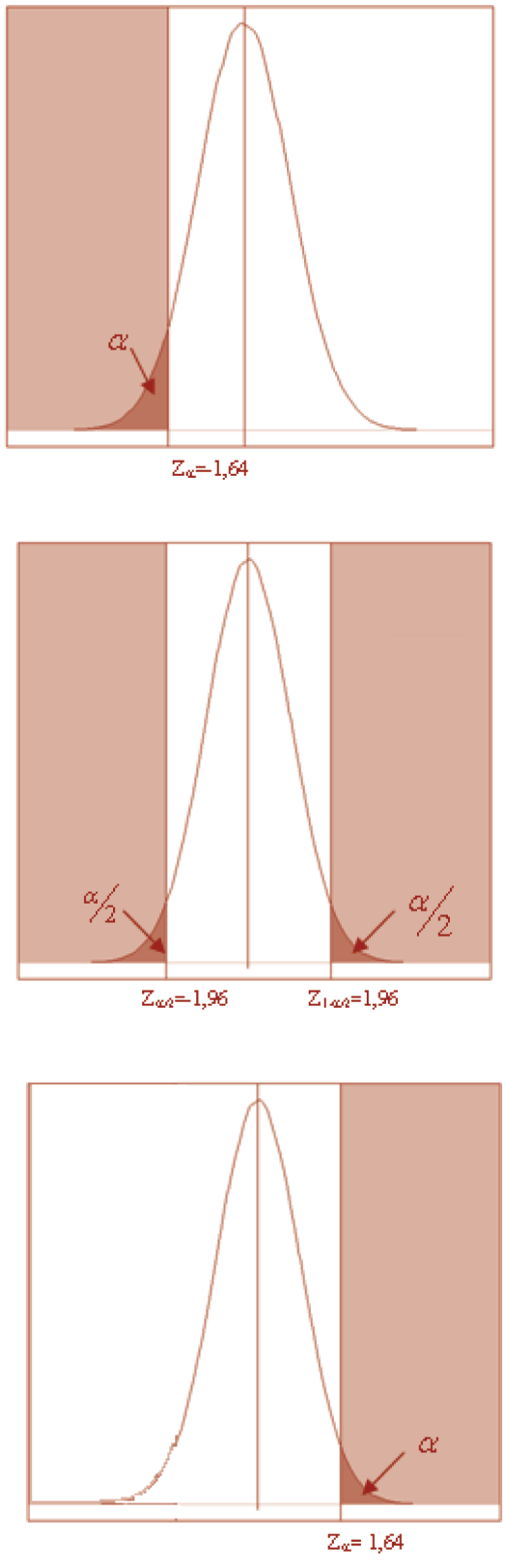

Figura 5. **Região crítica (a sombreado) para testes unilaterais à esquerda ou à direito e para um teste bilateral, considerado uma significância estatística de 5%**

exemplo, exige a utilização de testes bilaterais e a especificação deste facto na secção de métodos e no resumo dos artigos [6].

Exemplo

Para avaliar se a média de duas populações A e B são iguais a H_0 é a seguinte $H_0 : \mu_A = \mu_B$, ou, equivalentemente, $H_0 : \mu_A - \mu_B = 0$. Pode então escolher-se uma de três hipóteses alternativas, conforme o objectivo. Para testar se a média é diferente a um determinado parâmetro fixo, $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu > \mu_0$ (teste bilateral), se é maior que um determinado parâmetro fixo, $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu > \mu_0$ (teste unilateral à direita), ou se é menor que um determinado parâmetro fixo, $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu < \mu_0$ (teste unilateral à esquerda).)

CONCLUSÃO

A apresentação de resultados na mesma unidade das medições originais, juntamente com informação acerca da imprecisão inerente à variabilidade amostral, tem vantagens relativamente à apresentação apenas de valores P, geralmente na forma dicotómica "significativo" ou "não significativo".

A utilização de intervalos de confiança na literatura biomédica tem aumentado desde os anos 80, após a publicação do artigo intitulado *A show of confidence*, por Kenneth Rothman [7] e como resultado dos esforços editoriais que se lhe seguiram. Os *Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication* recomendam: "When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important information about effect size." [8].

Apesar da sua utilização quantitativamente crescente, os intervalos de confiança são frequentemente interpretados como se de testes de hipóteses se tratassem, ignorando a informação que encerram relativamente à precisão das estimativas [9]. Por outro lado, aspectos como a representatividade das amostras avaliadas ou a relevância clínica dos achados estatisticamente significativos são muitas vezes ignorados.

A compreensão destas ferramentas para inferência estatística, para além da simples leitura acrítica de *outputs* dos programas estatísticos da moda traduz-se em diferentes formas de comunicar mas sobretudo de compreender publicações biomédicas.

REFERÊNCIAS

1. Murteira BJF. Probabilidades e estatística, Volume II 2ª ed. Lisboa: Mcgraw-Hill; 1990.
2. Swinscow TDV, Campbell MJ. Statistics at square one, 10th ed. BMJ Books; 2002.
3. Altman DG, Machin D, Bryant TN, Gardner MJ, editors. Statistics with confidence, 2nd ed. BMJ Books; 2000.
4. Altman DG, editor. Practical statistics for medical research. London: Chapman &Hall; 1991.
5. Gardner MJ, Altman DG. *Confidence intervals rather than P values: estimation rather than hypothesis testing.* Br Med J (Clin Res Ed) 1986. 292(6522):746-50.
6. Instructions to authors. Journal of the National Cancer Institute [cited January 31, 2008]. Available from URL: http://www.oxfordjournals.org/our_journals/jnci/for_authors/index.html
7. Rothman KJ. A show of confidence. N Engl J Med. 1978;299(24):1362-3.
8. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication, Updated October 2007. International Committee of Medical Journal Editors. [cited January 31, 2008]. Available from URL: http://www.icmje.org
9. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. Psychol Sci 2004;15(2):119-26.

Correspondência:

HENRIQUE BARROS
Serviço de Higiene e Epidemiologia
Faculdade de Medicina da Universidade do Porto
Al. Prof. Hernâni Monteiro – 4200-319 Porto
Tel: +351 225513652 – Fax: +351 225513653
hbarros@med.up.pt